DOCUMENT RESUME

ED 290 322 FL 017 103

AUTHOR Meara, Paul; English, Fiona
TITLE Lexical Errors and Learners' Dictionaries.
PUB DATE 87
NOTE 12p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Dictionaries; Difficulty Level; *English (Second Language); Error Analysis (Language); *Error Patterns; *Instructional Effectiveness; Introductory Courses; *Media Selection; Second Language Learning

ABSTRACT

A study examined the effectiveness of English dictionaries in assisting in error correction by beginning learners of English as a second language. Lexical errors made in examinations were collected and coded by type, and the usefulness of the "Longman Active Study Dictionary", designed for ESL learners, in helping to correct the errors was analyzed. The results suggest systematic differences in error types made by students of different language backgrounds. It is proposed that (1) analysis of errors made by language learners can be useful to dictionary writers in increasing the materials' effectiveness and (2) a particular dictionary can vary in effectiveness among different target language groups. Further study is both recommended and planned. (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED290322

Dell

LEXICAL ERRORS AND LEARNER'S DICTIONARIES

Paul Meara
and
Fiona English

Birkbeck College
London University

address for correspondence:

Applied Linguistics Group
43 Gordon Sq
London WC1H OPD

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

P. Meara

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

FL017103

BEST COPY AVAILABLE

1 Background

The immediate stimulus for this report was a paper given at the 1986 meeting of the British Association for Applied Linguistics by Hilary Nesi. Nesi presented a study which suggested that dictionaries aimed at EFL learners were often ineffective, in that the entries provided either failed to prevent an obvious error, or actually reinforced an error by giving the impression that it was a correct use of a word. Clearly, if this claim is generalisable, then it has serious implications for dictionary writing. However, Nesi's work was based on a very small corpus of only 36 words, and something much more substantial would be needed to evaluate the claims she made. We were asked by the Longman Group to collect a larger corpus of errors with a view to making a similar analysis of dictionary effectiveness, but also to bear in mind the wider possibilities that such a corpus might offer as a general research tool.

2 The basic corpus

From December 1986 to January 1987 we collected and coded a substantial collection of lexical errors. The data consists of 1364 errors taken from a collection of First Certificate examination papers kindly provided by the Cambridge Examinations Syndicate. The errors came from a total of 14 language groups, but approximately 50% of the data was taken from essays produced by native speakers of Spanish.

We experimented with a number of different ways of handling this data, but eventually we coded each of the errors onto a database which recognised five different fields. These were:

a look-up word
a short context
a source language code
an error type code
a dictionary code

The **look-up word** is the word that the student would probably look up in the dictionary if he wanted to check what he had produced. i.e. in:

He was fond to drink a lot

the look-up word would be FOND, since users would reasonably expect to find the relevant information under the head adjective rather than under its accompanying preposition. The **short context** consists of as much context as necessary to make the error clear. **Language code** indicated the L1 of the learner who produced the scripts - this was usually inferable from the location of the Centre where the

examination took place.

The **error codes** were based on a simple 6 point system:

0 totally wrong word
1 phonologically related word
2 wrong word right semantic area
3 formal derivational errors
4 usage
5 spelling error

Code 0 covers cases where the error word is just wrong. Note that this code does not say why the error word is wrong, so that it treats errors arising from L1 transfer in exactly the same way as any other error of this type. This may seem slightly counterintuitive, but we justified on the grounds that the dictionary does not "know" what language background its user shares, and so cannot take this information into account.
Code 1 exists largely to deal with malapropisms and similar errors.
Code 2 covers words which are basically errors of style or register. The learner has chosen a word from the right semantic area which is not quite right. The reasons for this can vary.
Code 3 covers words which are essentially correct but wrongly constructed, e.g. REPETITIONAL instead of REPETITIOUS.
Code 4 covers cases where the word used is correct, but its context is incorrect, e.g. FOND TO instead of FOND OF.
Code 5 covers spelling errors where this results in a word other than the intended one. Other spelling errors are ignored. So, for example, LITE for LIGHT would not be included here, but LATE for LIGHT, or LIT for LIGHT would be.

This coding system is not entirely satisfactory in that there is some room for disagreement about which category any particular error belongs to, but on the whole, it worked rather better than a much more complicated system which attempted to make much finer distinctions. Examples of the error codes are provided in Table 1.

The **dictionary codes** codes indicate what would happen if the writer had used the Longman Active Study Dictionary, (a dictionary specifically aimed at EFL speakers), in order to check what s/he wanted to write. Bearing in mind that all the data is erroneous, there are three possible logical outcomes:

a) the dictionary identifies the error and shows the user how to correct it.
b) the dictionary identifies the error but fails to show the user how to correct it.
c) the dictionary fails to identify the error.

TABLE 1 Examples of error codes

0: completely wrong word:
  supply
  ...you would supply to change it...
  piles
  ...my tape-recorder had terrible piles...

1: phonologically related word:
  punch
  ...no sign of a punch on the tyre...

2: wrong word from the right semantic area:
  tranquilise
  ...my wife tranquilised me...

3: formal errors
  amuse
  ...there is an amusing arcade...

4: usage
  access
  ...he easily accessed drugs thanks to his money...

5: spelling (where this resulted in another word)
  prize
  ...the prize of the book was two pounds...

TABLE 2 Dictionary codes

C: clear: entry identifies the error and points to the correct word
Z: dead-end: entry identifies the error, but fails to offer any help
R: reference: as Z, but where reference to the correct word might have been expected
E: example: as Z, but where an example of correct usage would have prevented the error
P: prefix: as Z, but where additional information about other related forms would have prevented the error
X: invention: the word does not exist in the dictionary.
M: misleading: looking up the entry would not help to avoid the error

In practice, this three point system seemed rather too restricted, and we eventually adopted a more complex 7 point system which is exlained in more detail in Table 2.

Codes C and M correspond to cases (a) and (c) listed above. The remaining codes all correspond to particular instances of case (b). The paradigm of case (b) is Code Z, where the dictionary entry identifies the error, but fails to indicate what the writer should do about it. The other examples of this type, Codes R,E and P indicate cases where there is an obvious, systematic remedy for the failure of the dictionary to say how the error should be corrected.

3 Analysis of the data

The data has been analysed in two ways: a) the distribution of the errors on each of the codings separately, and b) interactions between pairs of codings.

3.a The individual codings

Figure 1a shows the basic distribution of errors according to error type. The largest component by far consists of semantically based errors (code 2), which account for almost half the entire corpus. The second largest component is made up of usage errors (code 4), which account for almost a quarter of the corpus. There are two other sizable components, formal errors (code 3) and completely wrong words (code 1), which each account for 15% of the errors. Phonologically based errors and spelling errors which resulted in an incorrect word together account for only 8% of the total corpus.

This distribution is quite interesting, and not entirely expected. In particular, the fact that semantically related errors, usage errors and formal errors together account for almost 80% of the total errors is especially important, since these are precisely the sorts of errors that dictionaries ought to be capable of preventing.

Figure 1b shows the distribution of the errors according to dictionary code. There are three important things to note in this data. Firstly, 33% of the entries are perfectly satisfactory (Code C). Secondly, only 6% of the entries are actually misleading (Code M). Thirdly, 21% of the entries are unsatisfactory, but could easily by improved by the addition of obvious extra information, (Codes R, E and P). This leaves a substantial number of entries which are unsatisfactory dead ends (Code Z). 35% of the entries fall into this category. (But cf. below where this conclusion is qualified by additional data.)

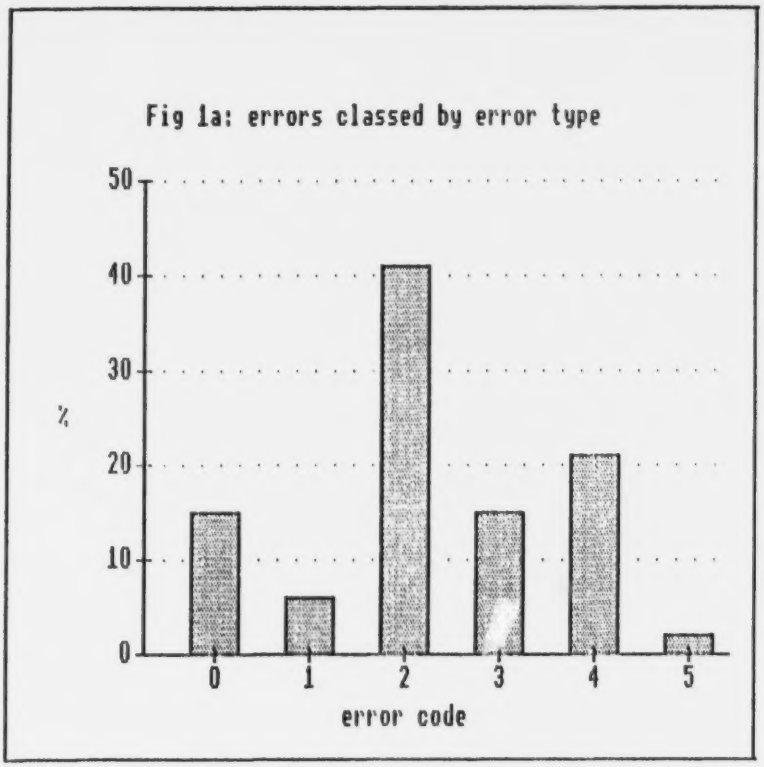
Fig 1a: errors classed by error type
50
40
30
%
20
10
0
0
1
2
3
4
5
error code

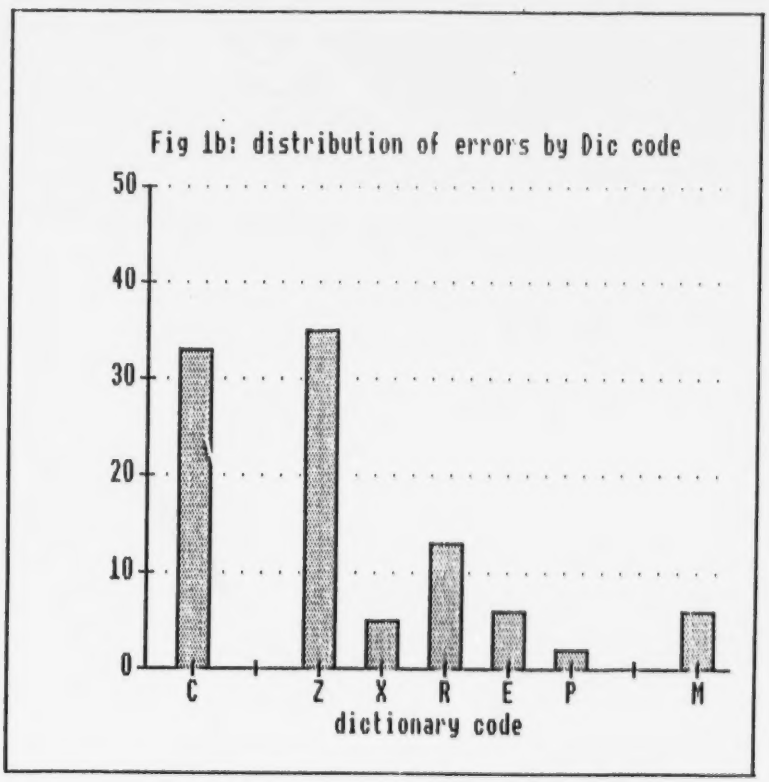
Fig 1b: distribution of errors by Dic code
50
40
30
20
10
0
C
Z
X
R
E
P
M
dictionary code

Since we have not compared this data with any other dictionary, it is not possible for us to say whether this distribution is good or bad. It is certainly rather better than the figures quoted by Nesi, and an informal, subjective assessment suggests that the Active Study Dictionary comes out rather well.

However, this analysis of the errors is a crude and superficial one, and a more complex picture emerges if we look at how these global figures break down under closer examination.

3.b Interactions between the codings

Three interactions will be reported in this section: the interaction between error types and dictionary codes; the interaction between error types and source language; and the interaction between source language and dictionary codes.

The interaction between error types and dictionary codes is reported in Table 3.

==========

TABLE 3:
Interactions between dictionary codes and error type
(small numbers of cases shown as - for simplicity)

d-code:	C	E	M	P	R	X	Z
e-type:							
0	-	-	-	-	-	26	168
1	-	-	-	-	-	24	53
2	101	-	78	-	134	-	233
3	136	-	-	22	25	-	-
4	189	75	-	-	-	-	12
5	-	-	-	-	-	-	-

==========

This table shows that our analysis of dictionary codes (Figure 1b) needs to be treated with considerable caution. The main point to emerge from this data is that dictionary codes are not distributed evenly among the various error types. Clear entries (Code C) are mainly associated with semantic errors, formal errors and usage errors. Dead ends (Code Z) are principally associated with wholly wrong words and with semantically related errors. This distinction is important, of course. Basically, there is no reason why one would expect a dictionary to be able to handle a wholly incorrect word, so that though the 168 Code Z entries found with type 0 errors are unfortunate from the user's point of view, they can hardly be treated as a fundamental inadequacy in the dictionary. This means that we

need to revise our estimate of the importance of dead ends in the Active Study Dictionary: if we ignore these 168 entries, then there are only 208 real dead ends - 15% of the whole corpus.

On the other hand, the 233 semantically related errors which evoked dead ends are obviously important. Two other important groups of errors also emerge from this analysis: the 78 semantically related errors that evoked misleading dictionary entries, and the 134 semantically related errors that could have been sorted if the dictionary had referred to an obviously related word. These three combinations suggest that if the dictionary paid more attention to the relationships between words in the same semantic area, its efficiency could be increased considerably. Solving all three of these problems would have doubled the number of CLEAR entries, and reduced the total of unsatisfactory entries to minimal levels.

The only other combination worth commenting on is the 75 errors which could have been avoided if an example of usage had been included in the dictionary entry.

The interaction between error types and source language is shown in Table 4. As in the previous data, this table has been simplified to make the patterns stand out more clearly.

This analysis reveals a number of interesting features. Basically, the proportion of different error types varies markedly from one language to another. Particularly noticeable is the high proportion of type 1 errors (phonologically related errors) for Chinese and Indonesian. In no other language does the proportion of such errors exceed 10%. Equally remarkable is the fact that Indonesian has a very small proportion of semantically related errors, whereas in all other languages in the sample this type of error accounts for at least 27% of the total. Less striking, but perhaps just as important are the variations in the other columns: some languages appear to have relatively small proportions of type 0 errors (completely wrong words) while for some languages these errors are relatively frequent; some languages give rise to relatively high proportions of type 3 errors (formal errors), while in other cases these formal errors are relatively few.

It is difficult to assess the importance of these distributions with any degree of certainty. The sample size for some languages is very small, and there may be a high degree of sampling error involved. The figures for Spanish will be fairly reliable, however, since the total sample size for that language is large (546 items). In this case, the bulk of the errors are semantic errors (42%), with a further 37% of the total evenly divided between formal errors and wholly wrong words. No other language produces a pattern of errors which resembles

this one very closely.

These findings are of some interest since they suggest that there might be systematic differences in the types of errors that learners from different language backgrounds produce. Most of the previous studies of errors have concentrated on errors collected from a single language group, or have compared two groups which are linguistically distinct but culturally similar - Finns and Swedes are the classical example of this. However, of even more interest is what emerges when we put together the data reported in Tables 3 and 4 and look at the interaction between source language and dictionary code. This data is reported in Table 5.

The best way of interpreting this table is to see it as an indication of how successful the Active Study Dictionary is with learners from different language backgrounds. A high figure in column C indicates that a high proportion of the errors made by that language group would have been picked up by the dictionary and corrected. This figure varies from a low of 21% to a high of 48%, indicating that the dictionary is more than twice as effective with some languages than it is with others. Equally disturbing is the variation in column Z, which represents dead-ends - cases where the dictionary identifies an error but fails to tell you what to do about it. This figure is best treated as a frustration index, perhaps: it varies from a low of 18% to a high of 61% : Swahili speakers are more than three times as likely to meet a dead end than are Finnish speakers speakers of about the same level. The surprising variations in the number of misleading entries (M) are also worth commenting on. In no case do they rise above 13%, but even this figure seems unacceptably high.

3. Discussion

When this data was presented at IATEFL, one member of the audience commented that 1364 errors could hardly be classed as a corpus. We accept this criticism, but nonetheless, it does appear that this work has thrown up a number of interesting ideas which deserve further study. Two ideas seem to be particularly important: a) the idea that an error corpus can provide sensible feedback to dictionary writers and suggest simple but efficient ways of increasing a dictionary's effectiveness; and b) the idea that a particular dictionary can vary in its effectiveness for different target language groups. We hope to be able to undertake a much larger project which should throw more light on these, and on other questions, in the near future. Our immediate plans are for a corpus of about 100,000 lexical errors collected from a rather narrower range of source languages than the fifteen reported here. A corpus of that order of magnitude could certainly not be dismissed as negligible, and should become an important research tool for the study of learners' lexical behaviour.

Paul Meara
Fiona English

Applied Linguistics Group
Birkbeck College London University

This research was supported by a grant from the Longman Group.

TABLE 4: Error types by source language
figures show percentage of errors in each category.
- indicates that less than 15% of errors fell in this category

e-type:	0	1	2	3	4	5
s-lang:						
French	-	-	40	-	27	-
German	-	-	51	-	19	-
Dutch	-	-	34	15	26	-
Spanish	18	-	42	-	19	-
Italian	27	-	49	-	-	-
Swedish	-	-	34	25	-	-
Norwegian	-	-	38	30	19	-
Finnish	-	-	31	31	22	-
Arabic	17	-	37	-	29	-
Japanese	-	-	49	-	27	-
Chinese	16	24	29	27	-	-
Greek	16	-	34	-	30	-
Indonesian	-	52	-	-	31	-
Swahili	25	-	27	16	16	-

TABLE 5: Dictionary Codes by Source Language
Figures show % of dictionary codings encountered learners from different language backgrounds.
- indicates that less than 10% of codings fell in this category.

d-code:	C	E	M	P	R	X	Z
s-lang:							
French	33	11	10	-	11	-	31
German	27	-	-	-	-	-	41
Dutch	34	-	10	-	10	-	28
Spanish	30	-	-	-	15	-	32
Italian	21	-	-	-	29	-	43
Swedish	34	-	-	-	15	-	28
Norwegian	34	-	11	-	15	-	23
Finnish	45	-	13	-	-	-	18
Arabic	44	-	-	-	13	-	31
Japanese	41	-	-	-	-	-	38
Chinese	25	-	-	-	-	-	52
Greek	48	-	-	-	10	-	24
Indonesian	48	12	12	-	-	-	24
Swahili	25	-	-	-	-	-	61